

WORD REORDERING APPROACH FOR MACHINE TRANSLATION FROM ENGLISH TO DRAVIDIAN LANGUAGES

J.Sangeetha

Department of CSE, Annamalai University, Annamalai Nagar, Chidambaram-608002.
jayasangita@yahoo.com

S.Jothilakshmi

Department of CSE, Annamalai University, Annamalai Nagar, Chidambaram-608002.
jothi.sekar@gmail.com

R.N.Devendrakumar

Department of CSE, SRIT, Coimbatore-641010
devapsna@gmail.com

ABSTRACT

In this proposed work, word re-ordering rules for English to Dravidian languages Machine Translation System is developed. Machine Translation (MT) mainly deals with the transformation of text from one natural language to another. Generally in Machine translation, the transformation hypothesis is computationally expensive. If we perform arbitrary reordering of words while translation, the search problem will be NP-hard. If we restrict the possible reordering in an appropriate way, we obtain a polynomial-time search algorithm. Word alignment differs for a sentence from one language to another language while perform machine translation. Re-ordering the given sentence improves the performance of the statistical translation. Here word re-ordering plays a very important role. In this paper, word reordering system for English to Dravidian languages is made to solve this problem in machine translation. The dependency information's are retrieved from Stanford parser incorporating with it; rules are created for the transformations. Transformations are then applied to English language to obtain the word order closer to the target Dravidian languages. This word reordering approach is preprocessing tool for machine translation which significantly improves the machine translation.

KEYWORDS:

Word reordering, reordering rules, SVO languages.

1. INTRODUCTION

Languages differ in the way that they order the words to produce sentences representing the same meaning. Source language input text need to reorder in order to produce fluent and equivalent output in the target language that retains the meaning of the source language text. Word reordering task plays a major role which significantly improves the performance of the machine

translation system. . In machine translation it has been shown that word reordering is a pre-processing step and it makes the translation process easier. The major structural difference between English and Indian languages is that English follows structure as Subject-Verb-Object, whereas, Indian languages follow default sentence structure as Subject-Object-Verb. Based on this, languages could be classified as SVO (English), SOV (Hindi, Tamil), VSO (Arabic), etc. With the differences of word order between English and Indian language, handling absolutely the reordering problem is very necessary.

Current phrase based machine translation systems can capture short range reordering via the phrase table. Even the capturing of these local reordering phenomena is constrained by the amount of training data available. For example, if adjectives precede nouns in the source language and follow nouns in the target language we still need to see a particular adjective noun pair in the parallel corpus to handle the reordering via the phrase table. Phrase based systems also rely on the target side language model to produce the right target side order. This is known to be inadequate (Al-Onaizan and Papineni, 2006), and this inadequacy has spurred various attempts to overcome the problem of handling differing word order in languages.

The weakness of these simple distortion models has been overcome using syntax of either the source or target sentence (Yamada and Knight, 2002; Galley et al., 2006; Liu et al., 2006; Zollmann and Venugopal, 2006). While these methods have shown to be useful in improving machine translation performance they generally involve joint parsing of the source and target language which is significantly more computationally expensive when compared to phrase based translation systems. Another approach that overcomes this weakness, is to reorder the source sentence based on rules applied on the source parse (either hand written or learned from data) both when training and testing (Collins et al., 2005; Genzel,2010; Visweswariah et al., 2010).

In this proposed work, word reordering system is made to improve the performance of the machine translation system. The dependency information's are retrieved from the Stanford parser incorporating with it; rules are framed for the transformations from English to Dravidian languages (SOV order languages). Transformations are then applied as a preprocessor to English language to obtain an underlying word order closer to the Dravidian languages.

The rest of the paper is organized as follows. Section 2 reviews related work and places our work in context. Section 3 outlines reordering issues due to syntactic differences between English and Tamil. Section 4 presents our reordering model, Section 5 presents experimental results and Section 6 presents our conclusions and possible future work.

2. RELATED WORK

There has been a lot of work on trying to improve the reordering model for a machine translation system. The original work on statistical machine translation was carried out by researchers at IBM. Their models are based on a string-to-string noisy channel model. The channel converts a sequence of words in one language (such as English) into another (such as French). The channel operations are movements, duplications, and translations, applied to each word independently. The movement is conditioned only on word classes and positions in the string, and the duplication and translation are conditioned only on the word identity. More recently, phrase-based models [2] have been proposed as a highly successful alternative to the IBM models.

Phrase-based models [3] generalize the original IBM models by allowing multiple words in one language to correspond to multiple words in another language. The automated transducer inference techniques OMEGA and GIATI work on phrase level but ignore the reordering problem from the view of the model. Without reordering both in training and during search, sentences can only be translated properly into a language with similar word order. In [6] weighted reordering has been applied to target sentences since defining a permutation model on the source side is impractical in combination with speech recognition. In order to reduce the computational complexity, this approach considers only a set of plausible reordering seen on training data. One criticism of the IBM-style translation model is that it does not model structural or syntactic aspects of the language. The model was only demonstrated for a structurally similar language.

The phrase based translation system [3] was a significant development in MT because the model was able to better estimate local reordering better than the IBM models [4], introduced a lexicalized block reordering model where two consecutive phrases can be swapped. The swapping of phrases allowed words to be reordered longer distances. [6] Proposed a distortion based model that allowed words to move any distance within the maximum distortion window. The distortion distance probabilities were computed using the word level alignment.

Beside the reordering methods during decoding, an alternative approach is to reorder the input source sentence to match the word order of the target sentence. Brown et al. (1992) discusses an analysis component for French which moves phrases around (in addition to other transformations) so the source and destination sentences are closer to each other in word order. There are approaches that targets translation of French phrases of the form NOUN1 de NOUN2. Method that combines morphologically-split verbs in German, and also reorders questions in English and German are also described. The reordering rules in their approach operate at the level of context-free rules in the parse tree. The results in their studies show that translation performance is significantly improved in BLEU score over baseline systems [2]. Michael Collins et al. [8] applied a sequence of handcrafted rules to reorder the German sentences in six reordering steps: verb initial, verb 2nd, move subject, particles, infinitives, negation. This approach successfully shows that adding syntactic knowledge can represent a statistically significant improvement from 1 to 2% BLEU score over baseline systems.

Our approach involves a preprocessing step, where sentences in the source language are to be translated are modified before being passed to an existing phrase based translation system. [10] We proposed a similar method for pre-processing the source language sentences for Indian languages. Their model employs the dependencies among the words in a sentence for reordering. But we use the context free rules extracted from the parsing information which governs the word order of English sentences.

3. ENGLISH-DRAVIDIAN LANGUGAESREORDERING ISSUES

This section provides similarity and differences of the two languages English and Tamil. The following are the divergences:

- English is a highly positional language with rudimentary morphology, and default sentence structure as SVO.
- Indian languages are highly inflectional, with a rich morphology, and default sentence structure as SOV.

- English uses prepositions while Tamil uses post-positions
- Dravidian languages allow greater word order freedom.
- Dravidian languages are relatively richer case-marking system

In addition, there are many stylistic differences. For example, it is common to see very long sentences in English, using abstract concepts as the subjects of sentences, and stringing several clauses. Such constructions are not natural in Indian languages, and present major difficulties in producing good translations.

As is recognized the world over, with the current state of art in MT, it is not possible to have Fully Automatic, High Quality, and General-Purpose Machine Translation. Practical systems need to handle ambiguity and the other complexities of natural language processing.

Hence to have a good translation system reordering the source sentence in accordance to target sentence is needed. And so reordering system for source sentences can make significant improvements over Indian languages.

4. REORDERING MODEL

The goal of transforming the source language input text into target Dravidian language word order is reordering. In the Machine Translation (MT) word reordering is serves as a preprocessing tool can help the machine translation process easier. Preprocessing tokenization, stemming and so on—is an essential step in natural language applications. Reordering of words on a sentence level as a more extensive step for preprocessing has succeeded in improving results in Machine Translation (MT).

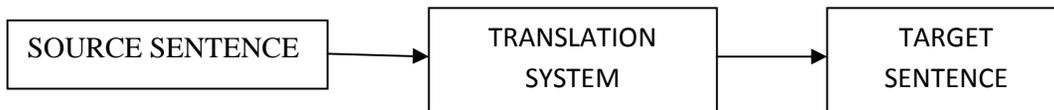


Fig. 1. General translation block diagram

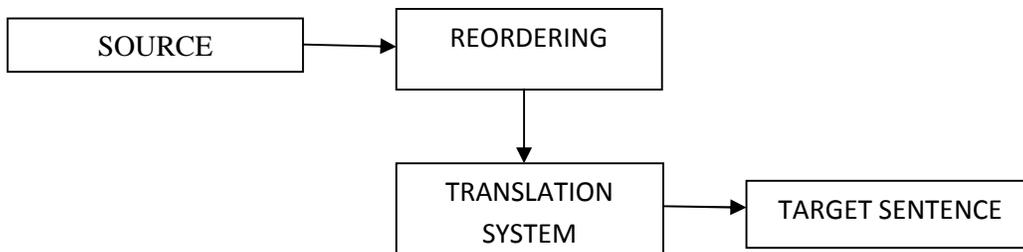


Fig. 2. Translation system with reordering

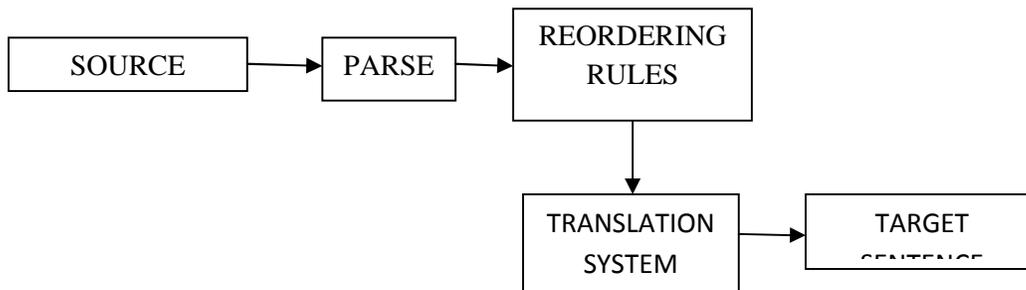


Fig. 3. Proposed reordering system

In figure 1 the general translation block diagram where the source sentence is given to the translation system and the output is produced. To improve the translation a system is designed with the reordering process. In Figure 2 the translation system is shown with reordering process. Here the Source sentence is given to the reordering system and then translated to the target sentence. The next figure shows how the reordering is done. In Figure 3 reordering block diagram is shown where the source languages are first parsed, for example using a Stanford parser it is parsed. A series of transformations is then applied to the resulting parse tree, with the goal of transforming the source language sentence into a word order that is closer to that of the target language is taken from the parse tree and the rule is generated.

Translation between SVO and SOV Languages:

In linguistics, it is possible to define a basic word order in terms of the verb (V) and its arguments, subject (S) and object (O). Among all six possible permutations, SVO and SOV are the most common. Therefore, translating between SVO and SOV languages is a very important area to study. The major structural difference between English and Indian languages is that English follows structure as Subject-Verb-Object, whereas, Indian languages follow default sentence structure as Subject-Object-Verb. Based on this, languages could be classified as SVO (English), SOV (Hindi, Tamil), VSO (Arabic), etc. With the differences of word order between English and Indian language, handling absolutely the reordering problem is very necessary. Consider the following sentence “He went to shop”.

Source sentence: He went to shop

Target sentence: Avan kadaikku sendran (Dravidian language -Tamil)

Where

He	:	Subject
Went to	:	Verb
Shop	:	Object

SVO in source sentence is transformed as SOV in target sentence

He	:	Subject	:	Avan
Shop	:	Object	:	Sendran
Went to	:	Verb	:	Kadaikku

In the above example the word order of the target sentence is not same as the word order of the source sentence. So after reordering the source sentence reordered form will be as

Source Sentence : He went to shop.
 Target Sentence :He shop went to.

This reordering is done according to human translation and done based on some specific rules. The rules are generated by getting the information from the parser. For the above sentence the general tree structure for the source side is shown in the figure 4figure 5 shows the rules for the target side how the sentence to be translated are hand written.

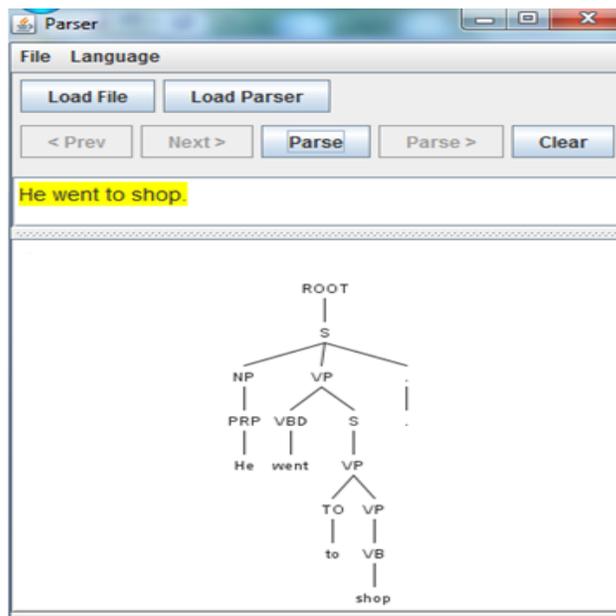


Fig. 4. General tree structure of source side

Depending on those rules every sentence pairs are translated. It is clear that in order for the phrase-based decoder to successfully carry out all of the reordering steps, a very strong reordering model is required. When the sentence gets longer with more complex structure, the number of words to move over during decoding can be quite high.

Sentence : He went to shop	
Source Side :	S -> NP VP VP -> VBD S
Target side:	S -> NP VP VP -> S VBD

Fig. 5. Reordering a simple sentence based on transfer rule

6. RESULTS AND DISCUSSION

The proposed technique has been implemented in Netbeans7.2. we are using Stanford parser to parse the English sentence to get its syntactic information. We are reordering the English sentence according to this structural information. To estimate the range of reordering performance (Kathik Viswaswariah et al 2011), consider the different POS bigrams in the input are reordered correctly. The proposed system has been implemented to reorder prepositions correctly, and to avoid any reordering that moves apart nouns and their adjectival pre-modifiers or components of compound nouns. Table 1 summarizes the reordering performance for these categories for a set of 50 sentences.). Each row in Table 1 indicates the total number of correct instances for the pair, i.e., the number of instances of the pair in the reference (column titled Total), the number of instances that already appear in the correct order in the input (column Input), and the number that are ordered correctly by the reordering model (column Reordered). The first two rows show that adjective-noun and noun-noun (compounds) are in most cases correctly retained in the original order by the model. The final row shows that while many prepositions have been moved into their correct positions.

7. CONCLUSION

The proposed method provides an effective methodology for reordering the source language as English sentences according to the target Dravidian languages. Reordering is important task which significantly improves the machine translation performance. Because different languages employ different word orders in their syntax, one requirement of an MT system is to get the target words in the right order. While phrase based MT systems do very well at reordering inside short windows of words, long-distance reordering seems to be a challenging task. Different researches have proven that preprocessing is the effective method in order to obtain a word-order which match with the word order of the target language. With this experiment we can prove that adding linguistic knowledge in preprocessing of training data can lead to remarkable improvements in translation performance. Moreover, we believe our approach can be generally applicable for other languages of which word orders are very different from English order.

Table 1: An analysis of reordering for a few POS bigrams

POS pair	Total	Input	Reordered
Adj-noun	100	82	85
Noun-noun	30	28	26
Prep-noun	150	50	25

Table 1: An analysis of reordering for a few POS bigrams

8. REFERENCES

- [1] R. Zens, H. Ney, T. Watanabe, and E. Sumita, "Reordering constraints for phrase-based statistical machine translation," Inspoken Language TranslationResearch Laboratories and Computer Science Department ATR and RWTH Aachen University and Germany Kyoto and japan.
- [2] http://en.wikipedia.org/wiki/Word_order.
- [3] Y. Al-Onaizan and K. Papineni, "Distortion models for statistical machine translation" In Proceedings of Association for Computational Linguistics,2006.
- [4] K. Yamada and K. Knight, "A decoder for syntax-based statistical machine translation" In Proceedings of Association for Computational Linguistics,2002.
- [5] M. Galley, J. Graehl, K. Knight, D. Marcu, S. DeNeefe, W. Wang, , and I. Thayer, "Scalable inference and training of context-rich syntactic translation models" In Proceedings of Association for Computational Linguistics, 2006.
- [6] Y. Liu, Q. Liu, , and S. Lin, "Tree-to-string alignment template for statistical machine translation" In Proceedings of Association for ComputationalLinguistics, 2006.
- [7] A. Zollmann and A. Venugopal, "Syntax augmented machine translation via chart parsing" In Proceedings on the Workshop on Statistical Machine Translation, 2006.
- [8] K. Yamada and K. Knight, "A syntax-based statistical translation model" InProceedings of the 39th Annual Meeting on Association for ComputationalLinguistics, 2001.
- [9] M. Collins, P. Koehn, and I. Kucerova, "Clause restructuring for statistical machine translation" In Proceedings on Association for ComputationalLinguistics, p. 531-540.
- [10] D. Genzel, "Automatically learning source-side reordering rules for large scale machine translation," In Proceedings of the 23rd International Conferenceon Computational Linguistics, 2010.
- [11] K. Visweswariah, J. Navratil, J. Sorensen, V. Chenthamarakshan, and N. Kambhatla, "Syntax based reordering with automatically derived rules forimproved statistical machine translation," In Proceedings of the 23rd International Conference on Computational Linguistics, 2010.
- [12] K. Visweswariah, R. Rajkumar, A. Gandhe, A. Ramanathan, and J. Navratil, "A word reordering model for improved machine translation," InProceedingsof the 2011 Conference on Empirical Methods in Natural Language Processing, 2011, pp. 486–496.

BIOGRAPHY NOTES ABOUT AUTHOR

J.Sangeetha received the B.E. degree in Computer science and Engineering from A.V.C college of Engineering at Mayiladurai in 2004. She received the M.E. degree in Computer Science and Engineering from Annamalai University in the year 2010. Doing her Ph.D. degree in Computer Science and Engineering at Annamalai University in the field of Speech to Speech translation from 2011. She published 3 papers in international journals and conferences. Her research interest includes speech processing, machine translation, speech recognition and speech synthesis.

Dr. S. Jothilakshmi received the B.E. degree in Electronics and Communication Engineering from Govt. College of Engineering, Salem in 1994. She received the M.E. degree in Computer Science and Engineering from Annamalai University in the year 2005. She has been with Annamalai University, since 1999. She completed her Ph.D. degree in Computer Science and Engineering at Annamalai University in 2011. She published 6 papers in international journals and conferences. Her research interest includes speech processing, image and video processing, and pattern classification.

R.N. Devendra Kumar received the B.E. degree in Computer Science & Engineering from P.S.N.A. College of Engineering & Technology, Dindigul in 2009. He received M.Tech. degree in Computational Engineering & Networking from Amrita University, Coimbatore in 2011.He has been working as Assistant Professor in Computer Science Department at Sri Ramakrishna Institute of Technology, Coimbatore. He is a Life Member in ISTE. His research includes computational linguistics, speech processing and machine translation.